

Reconciling Unevenly Sampled Paleoclimate Proxies: A Gaussian Kernel Correlation Multiproxy Reconstruction

J. L. Roberts^{1, 2*}, C. R. Tozer³, M. Ho⁴, A. S. Kiem⁵, T. R. Vance², L. M. Jong^{2, 6},
F. S. McCormack⁶, and T. D. van Ommen^{1, 2}

¹*Australian Antarctic Division, Kingston, Tasmania 7050, Australia*

²*Antarctic Climate & Ecosystems Cooperative Research Centre, University of Tasmania, Hobart, Tasmania 7001, Australia*

³*Oceans and Atmosphere, CSIRO, Hobart, Tasmania 7001, Australia*

⁴*Columbia Water Center, Columbia University, New York, New York 10027, USA*

⁵*Faculty of Science, University of Newcastle, Callaghan, New South Wales 2308, Australia*

⁶*Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania 7001, Australia*

Received 23 July 2017; revised 15 March 2018; accepted 21 June 2018; published online 19 August 2019

ABSTRACT. Reconstructing past hydroclimatic variability using climate-sensitive paleoclimate proxies provides context to our relatively short instrumental climate records and a baseline from which to assess the impacts of human-induced climate change. However, many approaches to reconstructing climate are limited in their ability to address sampling variability inherent in different climate proxies. We iteratively optimise an ensemble of possible reconstruction data series to maximise the Gaussian kernel correlation of Rehfeld et al. (2011) which reconciles differences in the temporal resolution of both the target variable and proxies or covariates. The reconstruction method is evaluated using synthetic data with different degrees of sampling variability and noise. Two examples using paleoclimate proxy records and a third using instrumental rainfall data with missing values are used to demonstrate the utility of the method. While the Gaussian kernel correlation method is relatively computationally expensive, it is shown to be robust under a range of data characteristics and will therefore be valuable in analyses seeking to employ multiple input proxies or covariates.

Keywords: climate, Gaussian kernel correlation, multiproxy, paleoclimate, reconstruction, uneven sampling

1. Introduction

The development of long, high-temporal resolution climate records (i.e., annual resolution spanning multiple centuries) from paleoclimate proxies is important for assessing low frequency variability, providing a context for recent climate extremes, and providing a baseline from which to assess impacts of human-induced climate change (e.g., Cook et al., 2010; McGregor et al., 2010; Gallant and Gergis, 2011; Ho et al., 2015b; Kiem et al., 2016). Paleoclimate proxies are the climate-sensitive physical, biological or chemical characteristics preserved in many natural archives spanning the world, such as ice cores, tree rings, corals, and sediments (e.g., McGregor et al., 2010; Batehup et al., 2015). Often these proxies represent similar climate processes or regimes, which facilitates the development of multiproxy reconstructions. For example, ice cores (Vance et al., 2013), corals (e.g., Evans et al., 2002; Cobb et al., 2003) and tree rings (e.g., D'Arrigo et al., 2005; Fowler et al., 2012) have all served as proxies of the El-Niño Southern Oscillation (ENSO).

Reconstructions based on multiple proxies are suggested to result in a more robust representation of target climate relative to single proxy reconstructions (e.g., Gergis and Fowler, 2009; Batehup et al., 2015); however, they are not without issues. These include the likely probability of multiple realisations of the same target climate from the different proxies e.g., inconsistency in the ENSO behaviour exhibited by different proxies (Wilson et al., 2010) and the (not unrelated) issue of varying sampling regimes of different proxies. We focus here on the latter issue as traditional linear regression-based reconstruction methods (e.g., composite plus scale) are not equipped to deal with unevenly and differently sampled proxies. Both the multi-proxy reconstruction methods of Li et al. (2010) and Hanhijärvi et al. (2013) allow for varying proxy resolution. However, both these methods require that the mapping of the proxies onto the same time-base as the target reconstruction be known a priori.

Here we present a new method for reconstructing climate using multiple proxy records based on the Gaussian kernel correlation method of Rehfeld et al. (2011). The Gaussian kernel correlation method has previously been applied to paleoclimate studies in several ways, for example, as a robust correlation estimate for assessing observations and climate model simulations of Asian rainfall (Rehfeld and Laepple, 2016), to

* Corresponding author. Tel: +61 3 6232 3179; fax: +61 3 6232 3179.
E-mail address: Jason.Roberts@aad.gov.au (J. L. Roberts).

estimate free parameters in the time-delay embedding method (Donges et al., 2015) and as the robust similarity estimator for the paleoclimate networks method of Rehfeld et al. (2013). In contrast, we apply an iterative optimisation technique to select an ensemble of possible paleoclimate reconstructions that match the Gaussian kernel correlation between a target series and (potentially multiple) proxy series. The approach uses a largely automated method of addressing differences in temporal resolution between the proxy records and the target variable as well as sampling irregularity within the proxy records. Robust uncertainty estimates are also generated. The method inherently weights each proxy independently from the temporal resolution of the proxy. The resulting reconstruction is scaled to produce a series with the same median and inter-quartile range as the target series over the calibration period used. The software is provided freely as supplementary material.

The following sections present a discussion of the method and examples of its implementation using synthetic and “real-world” climate proxy and gauged data.

2. Method

2.1. Method Overview

The method presented is based on generating climate reconstructions that match the existing Pearson correlation between each of the proxies and the target climate data series. A cost function (Equation 2), representing the mismatch between the actual correlation between the proxies and target climate and the correlation between the developed reconstruction and target climate, is minimised. It is assumed that the relationships between the target data series and the proxies are stationary, a necessary assumption for the development of preinstrumental climate reconstructions (e.g., Gallant et al., 2013). However, the method also allows for uncertainty in the independent variable (typically time) while optimising the reconstruction by minimising differences in the dependent variable.

The only knowledge about the target variable required for the reconstruction is its Pearson correlation with the proxies and quartiles for rescaling. The individual data points of the series are not required. For example, a reconstruction based on two proxies only requires knowledge of the proxies themselves and six additional statistics (i.e., the correlation, median and inter-quartile range for both proxies). The software provided here requires the target data series, but only for the calculation of these numbers using the same algorithms used to evaluate the potential reconstructions.

All proxies contribute equally to the reconstruction, regardless of the number of samples in the proxy. Therefore, the method inherently works with proxies of different temporal resolution. However, it should be noted that if a proxy is poorly correlated to the target, then its inclusion will do little to constrain the reconstruction. In addition, there are several constraints on the proxies. The proxy must be sampled at sufficient temporal resolution to ensure variability at the frequency of interest is captured by the proxies. In addition, the size of the

data gaps in the proxies and the frequency content of the proxies will determine which features can be successfully reconstructed. For example, attempting to reconstruct a monthly varying, strongly seasonal signal, across a data gap of several years will not successfully reproduce an annual cycle, let alone a seasonal cycle.

The Pearson method is linear, but it is possible to transform the proxies to linearise the relationship between the target and the proxy. The simplest such modification is a continuous piecewise linear transform. We provide separate code to help in the selection of suitable piecewise linear transformations. Further information about the methods is provided in Section 2.2 below.

2.2. Gaussian Kernel Pearson Correlation

The Gaussian kernel Pearson correlation (C_{GK}) between unevenly and differently sampled target series (t_i) and proxy series (p_j), of lengths n_t and n_p respectively, are calculated using the Gaussian kernel Pearson correlation slotting method of Rehfeld et al. (2011):

$$C_{GK}(t, p) = \frac{\sum_{i=1}^{n_t} \sum_{j=1}^{n_p} (t_i - \bar{t})(p_j - \bar{p}) K(d_i - d_{p_j})}{\sigma_t \sigma_p \sum_{i=1}^{n_t} \sum_{j=1}^{n_p} K(d_i - d_{p_j})} \quad (1)$$

where, \bar{t} and \bar{p} are the average of the two series t_i (target) and p_j (proxy), respectively, and d_t and d_p are the independent variables (typically time for proxy based reconstructions) for t and p respectively, and may differ from each other. Unlike Rehfeld et al. (2011) who normalise the signals to have zero mean and unit variance, we follow the method of Roberts et al. (2017), as this produces more robust estimates of the Pearson correlation. Specifically, we use the original signals t_i and p_j and correct for the mean and estimate the standard deviations (σ_t and σ_p) using the weighted summation Gaussian kernel ($K(d)$) from Equation 1. The Gaussian kernel correlation process inherently acts as a low pass smoothing filter, so the method is relatively insensitive to high frequency noise, with some increase in the uncertainty (see Figure 1). Not only can the distribution of samples differ, but the number of points (and the distribution of those points) for the reconstructed series (n_g) may be different from both n_t and n_p . The Gaussian kernel $K(d) = \exp(-d^2/2h^2)/(2\pi h)^{0.5}$ uses a width parameter (h). The selection of h influences somewhat the behaviour of the method, with larger values including more data at the expense of broadening the locality sampled. Unlike Rehfeld et al. (2011) and Roberts et al. (2017) who use a value of one quarter of the larger of the average spacing of the two data series for h , we use a value that is proportional (scaling constant h_s) to the maximum spacing between the data for either the target or proxy. This larger value of h ensures that the method is sensitive to points in data sparse regions and produces more robust solutions with smaller uncertainties. The choice of h_s is discussed more in Section 5.1, but in general is in the order of 0.25, and the method is robust to values of h_s near the optimum.

2.2.1. Correlation Matching the Target

From a randomly generated initial guess (g_k), of length n_g , as a possible reconstruction, we iteratively minimise the cost function (Equation 2):

$$CF = \sum_{i=1}^N \left(C_{GK}(t, p^i) - C_{GK}(g, p^i) \right)^2 \quad (2)$$

for the N proxies p^i .

The iterative optimisation is down gradient, where the gradient is estimated separately for each of the n_g points in g using the complex step method of Martins (2003). A down-slope correction is applied (with a new gradient calculated every iteration), until successively smaller corrections fail to improve the solution.

The variant of g_k , which produces the minimum cost function, produces a realisation of a reconstruction that is consistent with the target given its Pearson correlations with the proxies.

2.2.2. Reconstruction and Uncertainty

To produce a robust estimate of both the reconstruction and its uncertainty, an ensemble of 2000 solutions is used. Each solution is initialised to a randomly selected state (using a uniformly distributed random number sequence) and optimised to minimise the cost function. Each optimised ensemble member is offset and rescaled (neither altering the Pearson correlations with the proxies) to match the target series median and inter-quartile range respectively over the calibration period.

The reported reconstruction is the median of the 2000 ensemble members, again offset and rescaled to ensure matching of the target series over the calibration period. For robustness, the uncertainty (1 standard deviation) is calculated on a per-sample point basis as 1.483 times the inter-quartile range (Wilcox, 2010), where the factor compensates for the divergence in data coverage between the inter-quartile range and ± 1 standard deviation.

2.2.3. Implementation

The calculation of a robust estimation of the reconstruction, and associated uncertainties is computationally expensive. A naive implementation of Equation 1 would require, for each iteration of each ensemble member, the complex exponential to be calculated $n_g^2 n_p N$ times. Thus, for I iterations and 2000 ensemble members, a total of $2000 I N n_g^2 n_p$ complex exponential evaluations are calculated.

Several strategies are adopted to reduce this cost. Firstly, the proxies are invariant during calculation, and the independent variable (typically time) of the reconstructed series is also invariant. Therefore, we can precalculate all the exponential values once and store, and also precalculate all sums involving the proxies. Secondly, we optimise the maximum number of iterations (I), by testing the first eight ensemble members, successively doubling the number of iterations until conver-

gence. We select the maximum number of iterations required for these subsets of 8 ensemble members, applying an upper limit of 12800 iterations. Finally, we implement parallel computation using OPENMP code directives in the Fortran code to calculate multiple ensemble members concurrently. Together these strategies reduce the computational cost by several orders of magnitude, although remaining computational cost is not trivial. For example, the reconstruction described in Section 4.3, with 5 proxy records of millennial length and annual to sub-annual resolution being used to reconstruct a 1000-year annual resolved rain fall record, requires around 45 minutes on a quad core Intel i7 2.3 GHz laptop.

We provide a windows executable version of the code and source code for Fortran90, MATLAB, and Python in the supplementary information. We also provide code using the same Gaussian kernel algorithms to evaluate linearising the response of proxies via piecewise linear transformations (see Section 3.3 for more details).

3. Validation of Methods: Synthetic Examples

We use several synthetic datasets to evaluate the performance of both the method and the software implementation presented herein. Throughout these examples we use 2000 ensemble members (noting that small improvements in both the reconstruction and uncertainty are obtainable with an increased number of ensemble members, but the improvements are small compared to additional computational cost). We also limit the maximum number of iterations of the down slope solver to 12800, although only several hundred iterations are typically required.

3.1. Noise

Firstly, we investigate the performance when the target series has been contaminated with independent and identically distributed (*IID*) noise (η) uniformly distributed in the range $[-1, 1]$. In particular we consider the target as $t_i = \sin(d_i/10) + 0.4 \sin(d_i/3)$ for various noise levels (ϵ). The two proxies used are $\sin(d_i/10)$ and $\sin(d_i/3)$, and we use $h_s = 0.25$.

The accuracy of the zero-noise case (Figure 1a) is only limited by the maximum number of iterations (12800) for the down gradient solver. Increasing this upper limit reduces the errors, but at significant additional computational cost.

3.2. Missing Data

Next, we investigate the performance with missing data in the proxy data sets. We use the same target and proxies as in the noise test above, although with a larger value of $h = 0.5$ to reduce high frequency noise in the reconstructions (see Section 5.1 for more information). We consider three cases: where the lower frequency proxy is sampled at only 1/3 the rate of the other proxy (Figure 2b) and with 20% (Figure 2c) and 40% (Figure 2d) of the data is randomly and independently removed from both proxies. Finally, we consider the case of 20% missing data and $\epsilon = 0.4$ noise (Figure 2e).

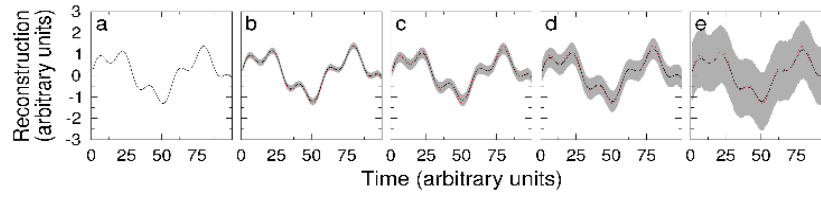


Figure 1. Target (red) and reconstruction (black) and 1 standard deviation range (gray bands) for the test signal $t_i = \sin(d_i/10) + 0.4\sin(d_i/3) + \varepsilon\eta$ for various noise levels, (a) $\varepsilon = 0$, (b) $\varepsilon = 0.1$, (c) $\varepsilon = 0.2$, (d) $\varepsilon = 0.4$, and (e) $\varepsilon = 1.0$.

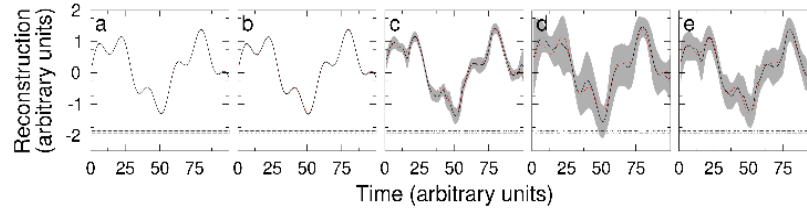


Figure 2. Target (red) and reconstruction (black) and 1 standard deviation range (gray bands) for the test signal $t_i = \sin(d_i/10) + 0.4\sin(d_i/3) + \varepsilon\eta$ for various missing data rates (data presence, but not value, indicated by gray squares at bottom of plot, lighter gray for lower frequency proxy), (a) no missing data, (b) $\sin(d_i/10)$ proxy sampled at 1/3 rate of other proxy, (c) 20% of data randomly and independently removed for each proxy, (d) 40% of data randomly and independently removed for each proxy, and (e) 20% missing data and $\varepsilon = 0.4$ noise.

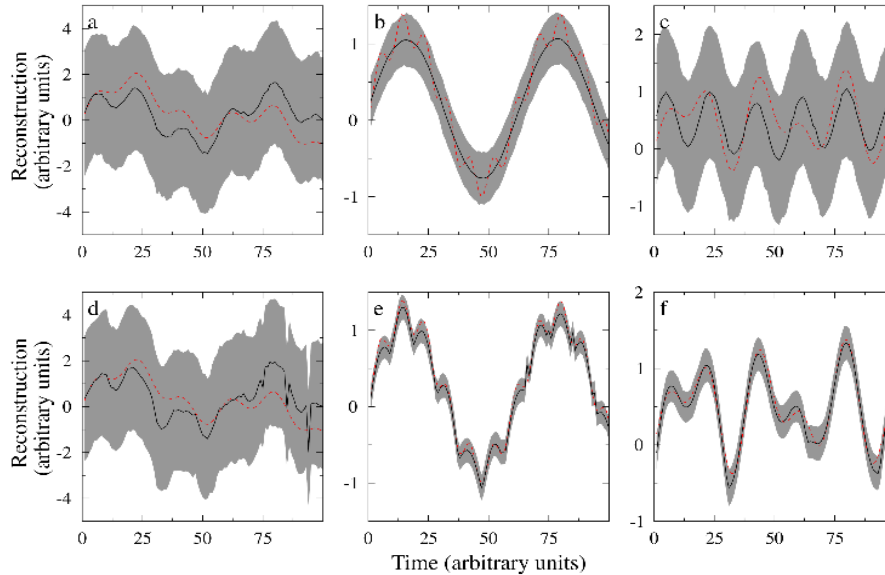


Figure 3. Target (red) and reconstruction (black) and 1 standard deviation range (gray bands). Upper row using original proxies and bottom row transformed proxies to enhance the Pearson correlation. (a) and (d) Unresolved target component of $\sin(d_i/20)$, (b) and (e) non-linear target $t_i = \sin(d_i/10) + 0.4\sin^2(d_i/3)$, and (c) and (f) non-linear target $t_i = \sin^2(d_i/10) + 0.4\sin(d_i/3)$.

3.3. Non-Linearity and Unresolved Components

Finally, we consider the cases of unresolved low frequency proxies and non-linearities, again defaulting to a value of $h_s = 0.25$. Performance with unresolved proxies is evaluated by the addition of a term $\sin(d_i/20)$ to the target without a corresponding proxy (Figures 3a and d). Performance in the presence of non-linearity is investigated using the same proxies but two revised targets of $t_i = \sin(d_i/10) + 0.4\sin^2(d_i/3)$ (Figures 3b and e), and $t_i = \sin^2(d_i/10) + 0.4\sin(d_i/3)$ (Figures 3c and f).

Performance in the presence of non-linearities can be improved by transforming the proxies to improve linearity. We include code to help with the selection of one possible transformation. In particular, we consider a continuous piecewise linear transformation (*PLT*), allowing up to a maximum of five points to define the remapping of the proxy. We optimise the *PLT* by maximising the quantity $\alpha C_{GK}(t, \hat{p})$, where \hat{p} is the *PLT* proxy, and the factor α compensates for the extra degrees of freedom associated with the number of points used to define

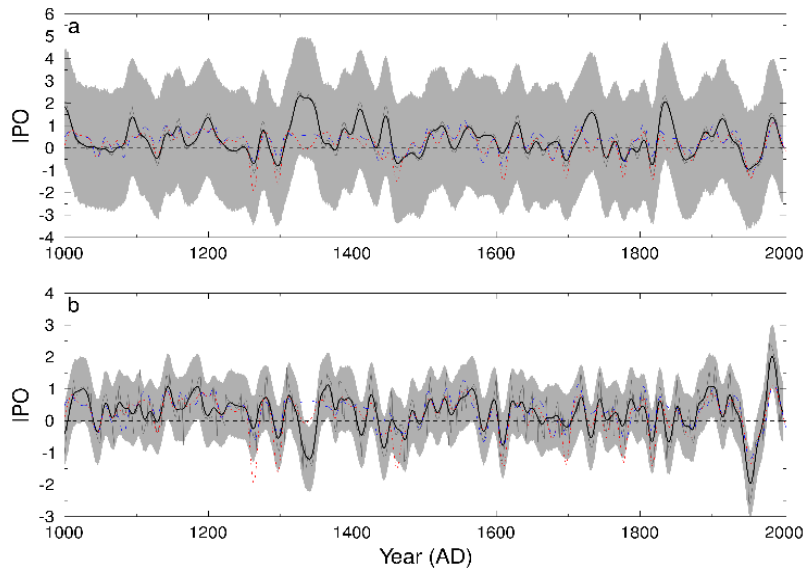


Figure 4. Reconstruction of the Interdecadal Pacific Oscillation (*IPO*) from Law Dome, East Antarctica, ice-core proxies. Median (gray line) and 1 standard deviation (gray bands) for 2000 ensemble members. (a) using unaltered proxies and (b) PLT proxies. Also shown are the two non-linear reconstructions of Vance et al. (2015), using a decision-tree (blue) and piece-wise linear (red) models. For direct comparison with Vance et al. (2015), we also show our reconstruction smoothed (black) with a Gaussian low-pass filter (equivalent half power width of 13 years).

the *PLT*. Specifically, similar to the generalised cross validation of Friedman (1991), we define $\alpha = (1 - 2N_{PLT} - 2/n_p)$, where N_{PLT} is the number of points defining the *PLT*. Again, the optimisation is via a down gradient scheme based on the complex step method (Martins, 2003).

If a meaningful improvement in Pearson correlation can be achieved by *PLT* compared to the original data series, then the quality of the regression can be significantly improved (compare Figures 3b and e, and c and f). Conversely, the indiscriminate application of the *PLT*, in cases where no significant improvement in the correlation is achieved, can degrade the quality of the reconstruction (compare Figures 3a and d).

4. Validation of Methods: Climate Variable Reconstruction

Here we present three examples of applications of the Gaussian kernel correlation method. An experimental reconstruction of the Interdecadal Pacific Oscillation (*IPO*) is made using three ice core-based proxy records, similar to Vance et al. (2015), to assess potential improvements in reconstruction skill. A second example is applied to infilling missing rainfall data for a gauge in the Lockyer Valley, *QLD*, using five other gauges located in the same basin. The third example is a demonstration of how precipitation in the eastern Murray Darling Basin (*) could be reconstructed using different paleoclimate proxy records.*

4.1. Reconstruction of the Interdecadal Pacific Oscillation

The Interdecadal Pacific Oscillation is a multi-decadal pat-

tern of sea surface temperature (*SST*) variability in the Pacific Ocean. Two complete positive (~1924 to 1943 and ~1979 to 1997) and one negative *IPO* phase (~1946 to 1976) have occurred in the instrumental period (Power et al., 1999; Verdon et al., 2004) and since the late-1990s the *IPO* is thought to be in its negative phase (Meehl et al., 2015). The positive *IPO* phase is associated with warming of the tropical Pacific and cooling of the north and south Pacific with the opposite *SST* patterns occurring during the negative phase (Power et al., 1999). *IPO*'s impact on Australia's rainfall variability is through the modulation of both the magnitude and frequency of *ENSO* impacts (Kiem et al., 2003; Kiem and Franks, 2004; Verdon et al., 2004; Power et al., 2006). Drought (flood) risk in eastern Australia is increased during *IPO* positive (negative) phases (Kiem et al., 2003; Kiem and Franks, 2004; Kiem and VerdonKidd, 2013) and therefore it is important to reconstruct past *IPO* conditions. This problem is challenging because it is both non-linear (Vance et al., 2015) and relatively large (n_t and $n_p > 1000$). Vance et al. (2015) reconstructed the *IPO* index of Parker et al. (2007) using three remote ice core proxies from Law Dome, East Antarctica: (i) annual snowfall, (ii) warm (*DJF-MAM*), and (iii) cold (*JJA-SON*) season sea salts. Vance et al. (2015) used two independent multivariate regression techniques, i.e., piecewise linear fit and decision tree, to reconstruct the *IPO*.

We use the method presented here to assess whether improvements to the reconstruction can be made and also as an independent verification of the reconstruction presented in Vance et al. (2015) (see Figure 4a for a reconstruction using the unaltered proxies and Figure 4b for the corresponding case using *PLT* proxies). We use a default value for $h_s = 0.25$ al-

though we obtain very similar results (percentage of explained variance for 13-year smoothed *PLT* case of 62.9 ~ 64.5%) for a range of $h_s = 0.01 \sim 2.0$.

To account for the non-linear response, we use the code we provide to evaluate possible piece-wise linear breakpoints in both the sea-salt based and snow accumulation rate proxies. The transformation of both proxies was optimal using a 5-point *PLT*, and increased the correlations with the instrumental *IPO* from 0.524 and 0.291 to 0.779 and 0.415 for the sea-salt and accumulation proxies respectively.

The relationship between the *IPO* and the proxies is known to be non-linear (Vance et al., 2015), so it is not surprising that the *PLT* proxies produce a higher quality reconstruction (62.9% explained variance for the instrumental *IPO* compared to 42.8% using the unaltered proxies). This is confirmed by the better agreement between the *PLT* reconstruction and the reconstructions of Vance et al. (2015) in Figure 4b. We observe all of the same key features in the *IPO* reconstruction as Vance et al. (2015), importantly including the sustained positive period in the 12th century corresponding to an epoch of exceptionally arid conditions in eastern Australia (Vance et al., 2015).

4.2. Infilling of Missing Rainfall Data

This example shows the use of the method to infill missing data, common in both proxy and gauged climate data. Six rainfall gauges (Table 1) were selected in the Lockyer Valley catchment in Queensland, Australia (Figure 5). The region receives the majority of its annual rainfall in the austral summer (December ~ February) months of the year.



Figure 5. Location of Lockyer Valley catchment (yellow) in southeast Queensland.

The rainfall at the Australian Bureau of Meteorology (*BOM*) high quality rainfall station (40082) is reconstructed using the remaining five rainfall records, all of which have periods of missing data. The reconstruction is for the period 1916 ~ 2015 with 10% and 50% of the calibration target data

removed. The 40082 rainfall record is successfully reconstructed in both cases (Figure 6), with little loss in fidelity (reduction of error $RE > 0.92$). Even with 50% of the data removed, a correlation with the original target data is $r = 0.96$ in both cases (i.e., more than 92% explained variance). This result is obtained using a value of h_s of 0.03, but is robust to a range of h_s , with in excess of 90% explained variance for h_s in the range $0.01 \sim 0.12$. The robustness of the reconstruction is evident by the agreement between the two reconstructions with 10 and 50% missing data, with a Pearson correlation of $r = 0.989$.

The scaling of the reconstructions is designed to reproduce the median and inter-quartile range of the target. However, other linear scalings of the data are possible, and such linear scalings will not change the correlations (see Section 5.2 for a more detailed discussion). For the two reconstructions, the median absolute deviation in monthly rainfall is 7.8 (6.5) and 10.8 (8.0) mm month⁻¹ for 10 and 50% data removal respectively, where the values in parentheses represent a linear rescaling to optimise with respect to median absolute deviation. Similarly, the root mean square errors (*RMSE*) are 18.0 (17.6) and 22.1 (20.3) mm/month.

Table 1. Details of the Lockyer Valley Rainfall Gauges

Site #	Site name	Epoch (AD)	Latitude (° S)	Longitude (° E)	Elevation (m)
40056	Coominya Post Office	1916–	27.39	152.50	81
40079	Forest Hill	1894–	27.58	152.38	112
40082	Gatton	1897–	27.54	152.34	89
40083	Gatton Allan St	1894–	27.54	152.28	114
40095	Hattonvale Oshea Rd	1908–	27.57	152.47	118
40424	West Haldon	1915–	27.76	152.08	336

4.3. Multiproxy Reconstruction of Rainfall in the Murray-Darling Basin

Here we use the method with two proxies with differing sampling regimes and temporal resolutions to reconstruct rainfall from a case study location in the Murray-Darling Basin. Specifically, we use the speleothem chemistry results from McDonald (2005) and McDonald et al. (2009) from Wombeyan Caves and the summer sea-salt record from Law Dome, East Antarctica of Vance et al. (2013) to reconstruct the *BOM* rain gauge record from Taralga Post Office (70080) for the instrumental period and last millennium. Taralga is located on the eastern border of the Murray-Darling Basin and precipitation near this station contributes to streamflow in the Lachlan River catchment. The region is largely dependent on dryland farming with minimal irrigation infrastructure. As a result, significant decreases in precipitation would likely adversely impact the regional economy, which is largely dependent on agriculture, and would likely damage regional wetlands and associated ecosystems.

For the calibration epoch (1884 ~ 2002, excluding years 1893 ~ 1894, which are missing from the rainfall record), the speleothem chemistry records for phosphorus (*P*), strontium (*Sr*), barium (*Ba*), and yttrium (*Y*) have Pearson correlations

with the gauged rainfall of -0.27 , 0.36 , 0.39 , and -0.28 respectively and are all significant at the 95% level (calculated using the method of Roberts et al. (2017)). It is noted that these correlations are opposite to the expected relationship between moisture and each of these species (Ho et al., 2015b). This could relate to the use of different calibration periods (and hence potential non-stationarity in the moisture-geochemistry relationship), lags in signal or assumptions in processing, but further evaluation of the geochemistry is beyond the scope of this paper. The Law Dome summer sea-salt record is correlated with the rainfall record at 0.23 and again is significant at the 95% level. The resulting reconstruction including the Law Dome data has a Pearson correlation of 0.48 (23% explained variance) compared to a best result of 0.36 (13% explained variance) excluding the Law Dome summer sea-salt record. This is obtained for $h_s = 0.09$ although again the result is robust to a range of h_s with more than 19% explained variance for h in the range $0.06 \sim 0.25$.

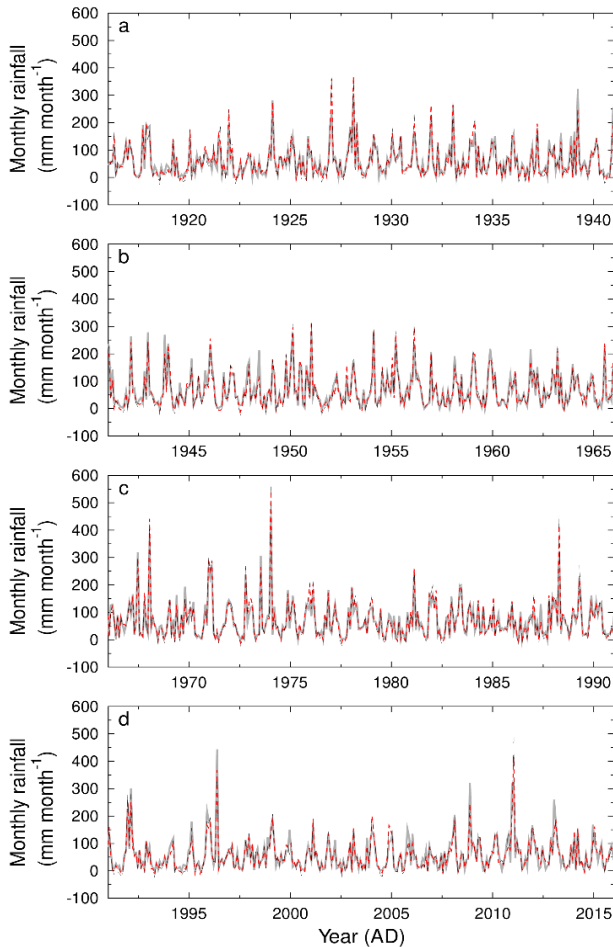


Figure 6. Rainfall reconstruction for the high-quality rainfall station 40082 for the period 1916 ~ 2015. Shown are the complete gauge rainfall record (gray), the reconstruction with 10% of the target data missing (red), and with 50% of the target data missing (black). Note that in many places it is hard to distinguish between the datasets due to the fidelity of the reconstruction.

The reconstruction for the last millennium is shown in Figure 7 and shows several interesting features, including a relatively wet epoch during the 20th century (when most of the instrumental records are available), and a very dry epoch during the 12th and early 13th centuries. This latter result is consistent with the south-east Queensland reconstruction of Vance et al. (2015). The high rainfall in the 20th century is consistent with the instrumental record (see Figure 8), and the reconstruction is robust to the removal of 50% of the calibration data (Figure 9).

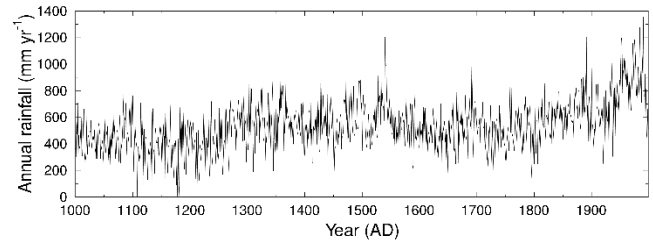


Figure 7. Rainfall reconstruction for Taralga Post Office based on Wombeyan Care speleothem chemistry and Law Dome summer sea-salt data.

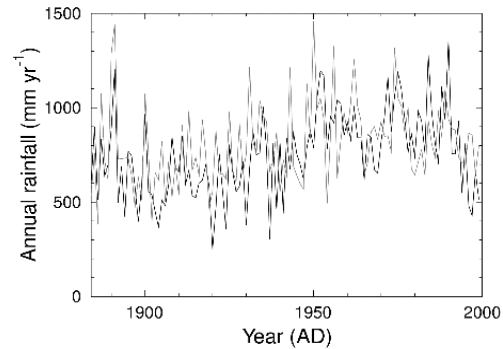


Figure 8. Rainfall reconstruction (black) and instrumental rainfall record for Taralga Post Office (gray). Note that in many places it is hard to distinguish between the datasets due to the fidelity of the reconstruction ($r^2 = 0.23$).

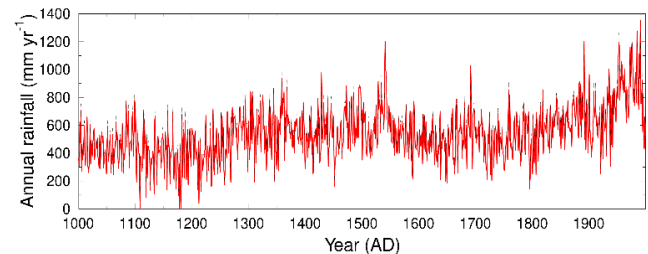


Figure 9. Rainfall reconstruction for Taralga Post Office using the entire calibration dataset (red) and with 50% of the calibration data removed (black).

The reconstruction here suggests that the long term climate is significantly drier than the 20th century, which raises questions about how representative the instrumental record is of longer term hydro-climate in this region, with obvious

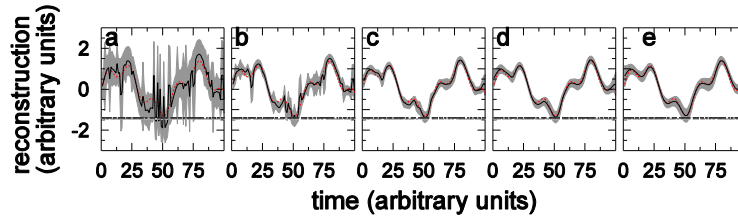


Figure 10. Target (red) and reconstruction (black) and 1 standard deviation range (gray bands) for the test signal $t_i = \sin(d_i/10) + 0.4\sin(d_i/3) + \varepsilon\eta$ with $\varepsilon = 0.4$ noise and 20% missing data (data presence, but not value, indicated by gray squares at bottom of plot, lighter gray for lower frequency proxy). A range of h_s is used for the reconstructions, (a) $h_s = 0.1$, (b) $h_s = 0.25$, (c) $h_s = 0.5$, (d) $h_s = 0.75$, and (e) $h_s = 1$.

consequences for water security policy based on assumptions of the representativeness of the instrumental record. However, this reconstruction is based on two proxies, one regional proxy with uncertainties in dating and magnitude of proxy response over time and another based on long-distance teleconnections. There is therefore a need to verify the demonstrated rainfall reconstruction presented here through the addition of other regional proxies that may capture different aspects of spatial and temporal rainfall variability (Ho et al., 2014). The future occurrence of extended periods of decreased precipitation in the region, as suggested by the reconstructed rainfall, would require advanced regional preparation and policies to ensure resilience or a strategy to evolve economic activities to adapt to a drier climate.

5. Discussion and Conclusions

The use of the Gaussian kernel Pearson correlation (Equation 1) allows for the accurate calculation of correlation coefficients for pairs of data series with different sampling frequencies, uneven and different sampling, missing data and different number of samples. These properties are inherited by the reconstruction method. In addition, the Gaussian kernel introduces some high frequency smoothing into the process, with the amount of smoothing increasing directly with the width parameter (h_s). Therefore, the method is relatively insensitive to high frequency noise, but also may not effectively capture genuine high frequency signals. The selection of h_s can have a significant impact on the quality of the reconstruction. This is discussed in more detail below (Section 5.1).

The Pearson correlation assumes a linear relationship between the two data series. However, the reconstruction method can be made more general by transforming the proxies to increase how close to linear the relationship between the proxies and the target is. Such a transformation should only be used if it results in a meaningful increase in linearity, otherwise the quality of the reconstruction may suffer. Therefore, we have chosen not to apply such a transform automatically, but instead provide users with a separate tool to calculate a potential transformation. The users should then apply domain specific expert knowledge to assess the suitability of using the transformed proxy.

While this method has several strengths (e.g., the ability to generate multi proxy reconstructions from unevenly and dif-

ferently sampled data), it has several limitations. Even after extensive optimisation of the computer code, it is relatively computationally expensive, with several of the examples shown here taking of order 10 minutes to complete on a quad core Intel i7 2.3 GHz laptop. Climate processes are often non-linear and while the method can produce non-linear reconstructions, this is achieved through transforming the proxies. A genuine non-linear method may produce a higher quality reconstruction. As with any method dealing with time-series data, the sampling must be at high enough frequency to resolve the modes of interest. In general, this involves several samples per mode, and this requirement will increase due to uneven sampling and the smoothing introduced by the Gaussian kernel. Missing data will typically increase the uncertainty of the estimate, while having a much smaller impact on the median estimate. Finally, as with any reconstruction method, the method assumes stationarity between the proxy and target climate variables. We suggest that users of this method (and indeed any reconstruction method) investigate the stationarity of the relationship between the proxy and target climate, which may include an assessment of the circulation processes linking the regions of interest (e.g., Gallant et al. 2013).

5.1. Influence of Gaussian Correlation Width Parameter

The parameter h_s influences the quality of the solution, although the optimal solution is obtained for a moderately broad range of h_s , so it is not necessary to overly refine the solution based on h_s .

Small values of h_s tend to make the reconstruction more reliant on local features of the proxy data. This may lead to the presence of high frequency noise in the reconstruction, and sensitivity to missing data. Alternatively, large values of h_s tend to overly smooth the reconstruction and lose fine scale features. For example, the 20% missing data case of Figure 2c shows high frequency noise and excessive sensitivity to missing data for small h_s (Figure 10).

While automation in the selection of h_s is possible, we feel that this is inappropriate. The selection of h_s should be actively undertaken by the user, to understand how sensitive (or insensitive) the reconstruction is to the selected value of h_s . However, we can provide some guidance. Typical values of h_s resulting in good reconstructions are of order 0.25, and the range 0.01 ~ 2 would be a good starting point. One useful strategy is to slowly increase h_s until the correlations between the target

and proxy data sets (reported by the computer code provided) start to decrease.

We have chosen to use a single value of h_s for all proxy datasets in a multi-proxy reconstruction, and normalise by the length of the largest gap between data. It would be easy to modify the code to use a different h_s for each dataset.

5.2. Linear Scaling of the Reconstruction

The Pearson correlation is invariant to both the addition of constant offset or the multiplication by a constant scaling factor. Therefore, the reconstructions can be linearly rescaled as required.

We have chosen to scale the reconstructions to have the same median and inter-quartile range as the target dataset over the calibration epoch. Linear scaling to minimise the median absolute difference or *RMSE* is possible, with the latter being more akin to a least squared solution. However, both of these scalings (or any alternative reconstruction method where the primary cost function is the minimisation of either of these quantities) will produce a reconstruction with greatly reduced variability, as indicated in Figure 11. While it is still possible to get insights into the frequency and duration of wet/dry epochs (Kiem and Franks, 2004; Tozer et al., 2016) and relative differences in large scale spatial climate variability (Ho et al., 2017) using these scalings, they are not suitable for studies reliant on reconstructions with realistic variability. These include studies aiming to quantify hydrological risk and how it varies over time or the assessment of changes in magnitudes of annual extremes over time (e.g., Kiem et al., 2016; Johnson et al., 2016), which are important in water resources management (Ho et al., 2016). The reconstruction (and associated scaling) method presented here hence may prove useful in water resources management and planning where there is a clear need to incorporate paleoclimate information (Tozer et al., 2016).

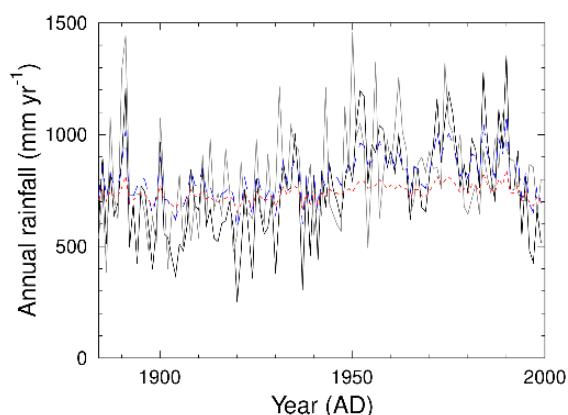


Figure 11. Linear rescaled reconstructions of rainfall for Taralga Post Office showing observed rainfall (gray), default (median and inter-quartile range) reconstruction (black), median absolute deviation rescaled (red) and *RMSE* (blue). Note the lack of variability compared to the observed rainfall in the later two rescaled reconstructions.

Acknowledgments. We thank Janece McDonald (Janece.McDonald@newcastle.edu.au) and Russell Drysdale (rnd@uni.melb.edu.au) for input relating to the Wombeyan Cave speleothem. See Ho et al. (2015a) and Ho et al. (2015b) for further details. The Australian Antarctic Division provided funding and logistical support (ASAC 757 and 4061). This work was supported by the Australian Government's Cooperative Research Centres Program-me through the Antarctic Climate & Ecosystems Cooperative Research Centre (ACECRC). This research was supported under Australian Research Council's Special Research Initiative for Antarctic Gateway Partnership (SR140300001). The Centre for Water, Climate and Land (CWCL) at the University of Newcastle provided partial funding for C. R. Tozer's salary.

References

- Batehup, R., McGregor, S., and Gallant, A.J.E. (2015). The influence of non-stationary teleconnections on palaeoclimate reconstructions of ENSO variance using a pseudoproxy framework. *Clim. Past*, 11 (12), 1733–1749. <https://doi.org/10.5194/cp-11-1733-2015>
- Cobb, K.M., Charles, C.D., Cheng, H., and Edwards, R.L. (2003). El Niño/Southern Oscillation and tropical Pacific climate during the last millennium. *Nature*, 424, 271–276. <https://doi.org/10.1038/nature01779>
- Cook, E.R., Seager, R., Heim, R.R., Vose, R.S., Herweijer, C., and Woodhouse, C. (2010). Megadroughts in North America: placing IPCC projections of hydroclimatic change in a long-term palaeoclimate context. *J. Quat. Sci.*, 25 (1), 48–61. <https://doi.org/10.1002/jqs.1303>
- D'Arrigo, R., Cook, E.R., Wilson, R.J., Allan, R., and Mann, M.E. (2005). On the variability of ENSO over the past six centuries. *Geophys. Res. Lett.*, 32 (3), 103711. <https://doi.org/10.1029/2004GL022055>
- Donges, J.F., Donner, R.V., Marwan, N., Breitenbach, S.F.M., Rehfeld, K., and Kurths, J. (2015). Non-linear regime shifts in Holocene Asian monsoon variability: potential impacts on cultural change and migratory patterns. *Clim. Past*, 11, 709–741. <https://doi.org/10.5194/cp-11-709-2015>
- Evans, M.N., Kaplan, A., and Cane, M.A. (2002). Pacific sea surface temperature field reconstruction from coral 180 data using reduced space objective analysis. *Paleoceanography*, 17 (1), 7–13. <http://doi.org/10.1029/2000PA000590>
- Fowler, A.M., Boswijk, G., Lorrey, A.M., Gergis, J., Pirie, M., McCloskey, S.P., Palmer, J.G., and Jan, W. (2012). Multi-centennial tree-ring record of ENSO-related activity in New Zealand. *Nature Clim. Change*, 2, 172–176. <https://doi.org/10.1038/nclimate1374>
- Friedman, J.H. (1991). Multivariate Adaptive Regression Splines. *Annals Stat.*, 19 (1), 1–67. <https://doi.org/10.1214/aos/1176347963>
- Gallant, A.J.E., Phipps, S.J., Karoly, D.J., Mullan, A.B., and Lorrey, A.M. (2013). Nonstationary Australasian teleconnections and implications for paleoclimate reconstructions. *J. Clim.*, 26, 8827–8849. <https://doi.org/10.1175/JCLI-D-12-00338.1>
- Gallant, A.J.E. and Gergis, J. (2011). An experimental streamflow reconstruction for the river Murray, Australia, 1783–1988. *Water Resour. Res.*, 47 (12), 158–166. <https://doi.org/10.1029/2010WR009832>
- Gergis, J.L. and Fowler, A.M. (2009). A history of ENSO events since A.D. 1525: implications for future climate change. *Clim. Change*, 92 (3), 343–387. <https://doi.org/10.1007/s10584-008-9476-z>
- Hanhijärvi, S., Tingley, M.P., and Korhola, A. (2013). Pairwise comparisons to reconstruct mean temperature in the Arctic Atlantic Region over the last 2,000 years. *Clim. Dyn.*, 41, 2039–2060. <https://doi.org/10.1007/s00382-013-1701-4>
- Ho, M., Kiem, A.S., and Verdon-Kidd, D.C. (2015a). A paleoclimate rainfall reconstruction in the Murray-Darling Basin (MDB), Australia: 1. evaluation of different paleoclimate archives, rainfall networks, and reconstruction techniques. *Water Resour. Res.*, 51 (10),

- 8362-8379. <https://doi.org/10.1002/2015WR017058>
- Ho, M., Kiem, A.S., and Verdon-Kidd, D.C. (2015b). A paleo-climate rainfall reconstruction in the Murray-Darling Basin (MDB), Australia: 2. assessing hydroclimatic risk using paleo-climate records of wet and dry epochs. *Water Resour. Res.*, 51 (10), 8380–8396. <https://doi.org/10.1002/2015WR017059>
- Ho, M., Lall, U., and Cook, E.R. (2016). Can a paleodrought record be used to reconstruct streamflow? A case study for the Missouri River basin. *Water Resour. Res.*, 52, 5195–5212. <https://doi.org/10.1002/2015WR018444>
- Ho, M., Lall, U., Sun, X., and Cook, E.R. (2017). Multiscale temporal variability and regional patterns in 555 years of conterminous U.S. streamflow. *Water Resour. Res.*, 53, 3047–3066. <https://doi.org/10.1002/2016WR019632>
- Ho, M., Verdon-Kidd, D., Kiem, A., and Drysdale, R. (2014). Broadening the spatial applicability of paleoclimate information - a case study for the Murray-Darling Basin, Australia. *J. Clim.*, 27, 2477–2495. <https://doi.org/10.1175/JCLI-D-13-00071.1>
- Johnson, F., White, C., van Dijk, A., Ekstrom, M., Evans, J., Jakob, D., Kiem, A., Leonard, M., Rouillard, A., and Westra, S. (2016). Natural hazards in Australia: floods. *Clim. Change*, 139 (1), 21–35. <https://doi.org/10.1007/s10584-016-1689-y>
- Kiem, A., Johnson, F., Westra, S., van, D.A., Evans, J., O'Donnell, A., Rouillard, A., Barr, C., Tyler, J., Thyer, M., Jakob, D., Woldemeskel, F., Sivakumar, B., and Mehrotra, R. (2016). Natural hazards in Australia: droughts. *Clim. Change*, 139 (1), 37–54. <https://doi.org/10.1007/s10584-016-1798-7>
- Kiem, A. and Verdon-Kidd, D. (2013). The importance of understanding drivers of hydroclimatic variability for robust flood risk planning in the coastal zone. *Aust. J. Water Resour.*, 17 (2), 126–134. <https://doi.org/10.7158/W13-015.2013.17.2>
- Kiem, A.S. and Franks, S.W. (2004). Multi-decadal variability of drought risk, eastern Australia. *Hydrol. Process.*, 18, 2039–2050. <https://doi.org/10.1002/hyp.1460>
- Kiem, A.S., Franks, S.W., and Kuczera, G. (2003). Multi-decadal variability of flood risk. *Geophys. Res. Lett.*, 30, 1035. <https://doi.org/10.1029/2002GL015992>
- Li, B., Nychka, D.W., and Ammann, C.M. (2010). The value of multiproxy reconstruction of past climate. *J. Am. Stat. Assoc.*, 105 (491), 883–895. <https://doi.org/10.1198/jasa.2010.ap09379>
- Martins, J.R.R.A. (2003). The complex-step derivative approximation. *ACM Trans. Math. Softw.*, 29 (3), 245–262. <https://doi.org/10.1145/838250.838251>
- McDonald, J. (2005). *Climate controls on trace element variability in cave drip waters and calcite: A modern study from two karst systems in S.E. Australia*. Ph.D. thesis, School of Environmental and Life Sciences, Newcastle, Australia.
- McDonald, J., Drysdale, R., Hodge, E., Hua, Q., Fischer, M., Treble, P.C., Greig, A., and J., H.C. (2009). One thousand years palaeohydrological record derived from SE Australian stalagmites. *EGU General Assembly Conference*, Eur. Geosci. Union, Vienna.
- McGregor, S., Timmermann, A., and Timm, O. (2010). A unified proxy for ENSO and PDO variability since 1650. *Clim. Past*, 6 (1), 1–17. <https://doi.org/10.5194/cp-6-1-2010>
- Meehl, G., Arblaster, J., and Chung, C. (2015). Disappearance of the southeast U.S. "warming hole" with the late 1990s transition of the Interdecadal Pacific Oscillation. *Geophys. Res. Lett.*, 42 (13), 5564–5570. <https://doi.org/10.1002/2015GL064586>
- Parker, D., Folland, C., Scaife, A., Knight, J., Colman, A., Baines, P., and Dong, B. (2007). Decadal to multidecadal variability and the climate change background. *J. Geophys. Res.*, 112 (18). <https://doi.org/10.1029/2007JD008411>
- Power, S., Haylock, M., Colman, R., and Wang, X. (2006). The predictability of interdecadal changes in ENSO activity and ENSO teleconnections. *J. Clim.*, 19 (19), 4755–4771. <https://doi.org/10.1175/JCLI3868.1>
- Power, S., Tseitin, F., Mehta, V., Lavery, B., Torok, S., and Holbrook, N. (1999). Decadal climate variability in Australia during the twentieth century. *Int. J. Climatol.*, 19 (2), 169–184. [https://doi.org/10.1002/\(SICI\)1097-0088\(199902\)19:2<169::AID-JOC356>3.0.CO;2-Y](https://doi.org/10.1002/(SICI)1097-0088(199902)19:2<169::AID-JOC356>3.0.CO;2-Y)
- Rehfeld, K. and Laepple, T. (2016). Warmer and wetter or warmer and dryer? Observed versus simulated covariability of Holocene temperature and rainfall in Asia. *Earth Planet. Sci. Lett.*, 436, 1–9. <https://doi.org/10.1016/j.epsl.2015.12.020>
- Rehfeld, K., Marwan, N., Heitzig, J., and Kurths, J. (2011). Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes Geophys.*, 18 (3), 389–404. <https://doi.org/10.5194/npg-18-389-2011>
- Rehfeld, K., Marwan, N., Breitenbach, S.F.M., and Kurths, J. (2013). Late Holocene Asian summer monsoon dynamics from small but complex networks of palaeoclimate data. *Clim. Dyn.*, 31, 3–19. <https://doi.org/10.1007/s00382-012-1448-3>
- Roberts, J., Curran, M., Poynter, S., Moy, A., van Ommen, T., Vance, T., Tozer, C., Graham, F., Young, D., Plummer, C., Pedro, J., Blankenship, D., and Siegert, M. (2017). Correlation confidence limits for unevenly sampled data. *Comput. Geosci.*, 104, 120–124. <https://doi.org/10.1016/j.cageo.2016.09.011>
- Tozer, C.R., Vance, T.R., Roberts, J.L., Kiem, A.S., Curran, M.A. J., and Moy, A.D. (2016). An ice core derived 1013-year catchment-scale annual rainfall reconstruction in subtropical eastern Australia. *Hydrol. Earth Syst. Sci.*, 20(5), 1703–1717. <https://doi.org/10.5194/hess-20-1703-2016>
- Vance, T., Roberts, J., Plummer, C., Kiem, A., and van Ommen, T. (2015). Interdecadal Pacific variability and Australian megadroughts over the last millennium. *Geophys. Res. Lett.*, 42, 129–137. <https://doi.org/10.1002/2014GL062447>
- Vance, T.R., van Ommen, T.D., Curran, M.A.J., Plummer, C.T., and Moy, A.D. (2013). A Millennial Proxy Record of ENSO and Eastern Australian Rainfall from the Law Dome Ice Core, East Antarctica. *J. Clim.*, 26 (3), 710–725. <https://doi.org/10.1175/JCLI-D-12-00003.1>
- Verdon, D., Wyatt, A., Kiem, A., and Franks, S. (2004). Multidecadal variability of rainfall and streamflow: Eastern Australia. *Water Resour. Res.*, 40(10), 333–341. <https://doi.org/10.1029/2004WR0032>
- Wilcox, R. (2010). *Fundamentals of Modern Statistical Methods*. Springer, New York, USA.
- Wilson, R., Cook, E., D'Arrigo, R., Riedwyl, N., Evans, M.N., Tudhope, A., and Allan, R. (2010). Reconstructing ENSO: the influence of method, proxy data, climate forcing and teleconnections. *J. Quat. Sci.*, 25 (1), 62–78. <https://doi.org/10.1002/jqs.1297>